Airway Gene-Expression Classifiers for Respiratory Syncytial Virus (RSV) Disease Severity in Infants

Lu Wang[1†], Chin-Yi Chu[2,5†], Matthew N. McCall[1], Christopher Slaunwhite[2], Jeanne Holden-Wiltse[1], Anthony Corbett[1], Ann R. Falsey[3,6], David J. Topham[4], Mary T. Caserta[5], Thomas J Mariani[2,5]*, Edward E. Walsh[3,6]* and Xing Qiu[1]*

[1]Department of Biostatistics and Computational Biology, [2]Department of Pediatrics and Division of Neonatology and Pediatric Molecular and Personalized Medicine Program, [3]Department of Medicine, [4]Department of Microbiology and Immunology, [5]Department of Pediatrics, University of Rochester School Medicine; and [6]Department of Medicine, Rochester General Hospital, Rochester, NY, USA

[†]These authors contributed equally to this work.

Corresponding authors:

Thomas J. Mariani
Email: Tom_Mariani@urmc.rochester.edu
Phone: 585-276-4616
Fax: 585-276-2643

Edward E. Walsh
Email: Edward.walsh@rochesterregional.org
Phone: 585-922-4331
Fax: 585-922-5168

Xing Qiu
Email: Xing_Qiu@urmc.rochester.edu
Phone: 585-275-0666
Fax: 585-273-1031

Author Contributions

XQ, TJM, and EEW conceptualized the study. TJM, EEW, MTC, and CC designed the experiments. EEW, MTC, ARF and DJT developed the cohort, and collected the specimens and clinical data. LW, MNM, and XQ developed statistical models. JHW and AC facilitated data organization, management and analysis. LW, CC, MNM, CS, JH-W, AC, ARF, DJT, MTC, TJM, EEW, and XQ generated, analyzed and interpreted the data. LW, CC, MNM, CS, JH-W, AC, ARF, DJT, MTC, TJM, EEW, and XQ wrote and/or revised the manuscript. All authors read and approve the final manuscript.

Funding Sources: This study is supported in part by Respiratory Pathogens Research Center (NIAID contract number HHSN272201200005C), and the University of Rochester CTSA award number UL1 TR002001 from the National Center for Advancing Translational Sciences of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Financial Disclosure: The authors have no financial relationships relevant to this article to disclose.

Conflict of Interest: The authors have no conflicts of interest relevant to this article to disclose.

Abstract word count: 238.

Text word count: 3037

Abstract

Background: RSV infection is common in infants, with a majority of those affected displaying mild clinical symptoms. However, a substantial number of infants infected with RSV develop severe symptoms requiring hospitalization. We currently lack sensitive and specific predictors to identify a majority of those who require hospitalization.

Method: We used our previously described methods to define comprehensive airway gene expression profiles from 106 full-tem previously healthy RSV infected subjects during acute infection (day 1-10 of illness; 106 samples), and during the convalescence stage (day 14-28 of illness; 69 samples). All subjects were assigned a cumulative clinical illness severity score (GRSS). High throughput RNA sequencing (RNAseq) defined airway gene expression patterns in RSV infected subjects. Using AIC-based model selection, we built a sparse linear predictor of GRSS based on the expression of 41 genes (NGSS1). Using a similar statistical approach, we built an alternate predictor based upon 13 genes displaying stable expression over time (NGSS2). We evaluated predictive performance of both models using leave-one-out cross-validation analyses.

Results: NGSS1 is strongly correlated with the disease severity, demonstrating a cross-validated correlation (ρ) of 0.813. As a binary classifier (mild versus severe), NGSS1 correctly classifies 89.6% of the subjects following cross-validation. NGSS2 has slightly less, but comparable, prediction accuracy with a cross-validated correlation of 0.741 and cross-validated classification accuracy of 84.0%.

Conclusion: Airway gene expression patterns, obtained following a minimally-invasive procedure, have potential utility as prognostic biomarkers for severe infant RSV infections.

Introduction

Respiratory Syncytial Virus (RSV) is the most important cause of respiratory illness in infants and young children, accounting for more than 100,000 bronchiolitis and pneumonia hospitalizations in the US annually.[1] Worldwide, 33.1 million acute lower respiratory infections and 3.2 million hospitalizations in children under 5 years of age are attributed to RSV each year.[2] In the US ~1-2% of newborns are hospitalized during their first winter, with rates greatest in the first month of life (25.9 per 1000).[3] Risk factors for severe disease include gestational age < 29 weeks, bronchopulmonary disease and symptomatic congenital cardiac disease, while less well defined risks include lack of breast feeding, and exposure to tobacco smoke. However, the majority of hospitalized infants are full-term infants whose only risk factor is young age at the time of infection.[3]

A number of severity scores using clinical parameters, including cutaneous oximetry, have been proposed for quantifying illness severity and to guide decisions to hospitalize RSV-infected infants.[4-13] However, none of the clinically based severity scores have been universally adopted.[14] Reasons may include heterogeneity in the scope and purpose of the score, the ages to which it is applied and concerns about inter-observer variability and subjectivity in interpreting clinical signs, including oximetry, that often are temporally dynamic over short intervals. Finally, the predictive accuracy of clinical severity scores has not been clearly demonstrated. Identification of an objective biomarker that accurately correlates with, or potentially predicts, disease severity could be highly useful in the clinical assessment and management of RSV-infected infants.

We and others have reported a relationship between disease severity and host gene expression in peripheral blood cells, and nasal swab samples, during infection.[15-18]

These results suggest such an approach may allow development of predictive biomarkers to accurately categorize severity and potentially differentiate more vulnerable subjects from those not at risk of severe illness. As part of the AsPIRES study of RSV pathogenesis in healthy full-term infants with primary RSV infection.[19] we recently reported on the feasibility of measuring gene expression of airway cells collected by nasal swab prior to and during RSV infection.[20] In this manuscript, we describe the development of two airway gene expression-based classifiers that are highly correlated with clinical disease severity. This represents a major step in developing a gene expression diagnostic capable of classifying clinical severity in RSV-infected infants.

Methods

Study Subjects: Subjects included RSV infected infants enrolled in the AsPIRES study at the University of Rochester Medical Center (URMC) and Rochester General Hospital (RGH).[19] The Institutional Review Boards of both institutions approved the study, and all parents provided written informed consent. RSV-infected infants came from three cohorts during three winters (October 2012 through April 2015); one cohort included infants hospitalized with RSV, a second cohort was recruited at birth and followed through their first winter for development of RSV infection, and the third cohort was RSV infected infants seen in pediatric offices and emergency departments and managed as outpatients. All subjects were full-term infants undergoing a primary RSV infection during their first winter season. Nasal samples were collected from the inferior nasal turbinate, by gentle brushing with a flocked swab as previously described [20], during the acute illness visit (visit 1) and at a convalescent visit ~28 after illness onset (visit 2). Illness severity was graded from 0 to 10 using a Global Respiratory Severity Score (GRSS), as previously described.[21]

Nasal RNA processing: The process for nasal RNA recovery was previously described.[20] Briefly, following flushing of the nares with saline to remove mucus and cellular debris, a flocked swab was used to recover cells at the level of the turbinates. The swab was immediately placed in RNA stabilizer (RNAprotect, Qiagen, Germantown, MD) and stored at 4 °C. Cells were recovered by filtering through a 0.45 uM membrane filter. Recovered cells were lysed and homogenized using the AbsolutelyRNA Miniprep kit (Agilent, Santa Clara, CA) according to the manufacturer's instructions. 1 ng of total RNA was amplified using the SMARter Ultra Low amplification kit (Clontech, Mountain View, CA) and libraries were constructed using the NexteraXT library kit (Illumina, San Diego, CA). Libraries were sequenced on the Illumina HiSeq2500. Sequences were aligned against human genome version of hg19 using STARv2.5, counted with HTSeq, and normalized by Fragments Per Kilobase of transcript per Million mapped reads (FPKM).

RNAseq data preprocessing

The nasal RNA samples generated an average of 27 ± 6 million total reads, with 66 ± 14% displaying unique mapping to the human genome. Fifteen samples had low yields (< 5 million mapped reads) and were excluded from analyses. Some samples (n=10) with low pair-wise correlation with the other samples (mean $\rho < 0.43$) were also removed prior to analysis. We filtered the data set to exclude genes with very low normalized expression levels across all subjects. After these quality assurance analyses, the data set consisted of expression data for a total of 13,819 genes in 175 samples. Because the study spanned three winter seasons, samples were processed in six library batches (Supplementary Figure E1). We applied ComBat [22] to remove batch effects prior to statistical analyses. There also were a small proportion of outliers with disproportionally large expression values that could skew the statistical analyses, thus we winsorized the top and bottom 1% of gene expressions. We also excluded from the

analysis 131 genes with more than half of their expression values winsorized. Finally, to ensure the robustness of the multivariate classifiers to be developed in this study, we filtered half of the remaining genes based on the inter-quantile range (IQR) criterion. In this way, the remaining 6,844 genes are guaranteed to have a relatively large range of expression values.

Statistical methods:

Descriptive statistics are reported in Table 1. Discrete variables are summarized in percentages, and continuous variables were summarized as Mean (SE). For continuous variables, we performed 2-sample Welch t-tests to check the equality between the mild and severe groups; for categorical variables, Fisher's exact test was used instead. The nasal gene-expression severity scores we developed in this study were based on multivariate regression analysis with bi-directional stepwise model selection based on Akaike Information Criterion (AIC). Technical details of model development and cross-validation, including a comparison between AIC-based approach (Strategy 1) and an alternative approach based on elastic-net regression [23, 24] (Strategy 2) can be found in Supplementary Material. All analyses were conducted using SAS 9.3 (SAS Institute Inc., Cary, NC, USA) and the R programming language (version 3.5, R Foundation for Statistical Computing, Vienna, Austria).

Results

Of the 139 RSV-infected infants enrolled in the AsPIRES study, nasal samples were available from 119 subjects during acute infection (day 1-10 of illness) and 81 subjects during convalescence (day 14-28 of illness). Among these 200 samples, 175 samples (106 acute samples and 69 convalescent samples) met sufficient quality to be used for subsequent analyses. Demographic and clinical information for these 106 subjects are provided in Table 1. The clinical severity score (GRSS) for these subjects ranged from 0

to 10, with 42 subjects considered to have mild disease (GRSS ≤3.5; mean ± SE GRSS of 1.63 ± 0.15) and 64 to have severe disease (mean GRSS of 6.13 ± 0.22). There were no significant differences between the mild and severe groups in gender, race, delivery type, breast feeding, or exposure to tobacco smoke. There also was no difference in age at time of infection or in duration of illness at the time of evaluation.

Nasal gene expression correlates of clinical severity during acute illness

The 6,844 genes remaining after data preprocessing and filtering were subjected to the Pearson correlation test to select genes that were significantly correlated with GRSS during acute infection. After applying the Benjamini-Hochberg procedure to control false discovery rate (FDR) at the 0.05 level, 66 significant genes (top correlates) were identified.[25] Using these 66 genes, we applied model selection procedures (see Supplementary Text for more details) to select an initial multivariate regression model for GRSS (Model 1), which was comprised of 39 genes. Using leave-one-out cross-validation, we showed that this initial model had a relatively good predictive power (77.4% prediction accuracy, or 24 misclassifications) for the dichotomous clinical outcome (mild vs. severe illness).

We performed a principal component analysis (PCA) based on the decomposition of the correlation matrix of the 66 genes most highly correlated with GRSS. We found that the first principal component (PC1) explained a very large proportion (46%) of total variance, with PC2 explaining only 7.4% of the variance. This suggested a strong correlation among the 66 genes, which might reduce the performance of Model 1. Therefore, we sought to identify supplementary genes, with low correlation with PC1, as additional features to model GRSS (see Supplementary Text for more details). We applied the same model selection procedure to the union of original 66 top correlates

and 10 supplementary genes (a total of 76 candidate features), and developed Model 2 that was comprised of 41 genes.

For comparison, we also performed a similar model selection procedure using the 76 genes most correlated with GRSS (original set of 66 top correlates plus the next 10 most significant genes). This Model 3 was comprised of 42 genes.

The performance of each of the three models was evaluated by leave-one-out cross-validation (Table 2). We found that the incorporation of the supplementary genes into Model 2 (cross-validated prediction accuracy of 89.6%; 11 misclassifications) significantly improved the accuracy compared to Model 1 (24 misclassifications) and Model 3 (23 misclassifications). Of note, Model 2 contained 5 of the supplementary genes, and we have defined this optimal model as NGSS1 (nasal gene expression severity score 1).

As shown in Figure 1, there is a high correlation between NGSS1 and the clinical severity score (GRSS) in the naïve (ρ=0.935) and cross-validated analyses (ρ=0.813). Furthermore, NGSS1 predicted a mean GRSS of 6.22 for severely ill infants (GRSS>3.5) during acute illness (Figure 2). We also calculated the NGSS1 at convalescence (~day 28 after illness onset) for the 54 subjects with data available from both acute and convalescent time points. Since most infants had recovered from their illness by this time, we expected that gene expression would be less informative for classifying disease severity. NGSS1 predicted a significantly lower mean GRSS (2.82, $p<.001$) for severely ill infants (GRSS>3.5) during convalescence. However, there was no difference in NGSS1 between visits for the mildly ill group (1.96 vs. 2.31, $p=0.45$).

Stable nasal gene expression correlates of clinical severity

We next developed a nasal gene expression severity score based on genes displaying stable expression across acute illness and convalescence in the 54 subjects with

samples from both time points. Specifically, we included only genes whose mean expression levels correlated with disease severity during acute illness, and whose expression did not change significantly from the acute to convalescent stage.
We identified 2127 genes in subjects with mild illness (GRSS ≤3.5) and 1531 genes in subjects with severe illness (GRSS >3.5), based on paired two sample t-test ($p > 0.5$) and fold change increases or decreases within 10%. Of the total 3658 genes, 689 stable genes were informative in both groups (Figure 3A). A quality assurance analysis based on IQR showed that a small subset (n=14) of these genes had relatively small dynamic range in the combined dataset, and were excluded. The remaining 675 stable genes were used in model development.

We applied marginal screening based on Pearson correlation with GRSS for all 106 subjects, and identified 44 marginally significant genes. As in developing NGSS1, we added 5 supplementary genes that are least correlated with them yet still have strong marginal associations with the GRSS. We repeated our regression-based/step-wise model selection procedure and identified 13 genes as Model 4 (also designated as **NGSS2**). The performance of NGSS2 is provided in Table 2 and illustrated in Figure 3B. NGSS2 showed a strong and significant correlation with GRSS (ρ=0.741), and a cross-validated accuracy of 84% (17 misclassifications, Table 2). Of note, NGSS1 and NGSS2 do not contain any commonly selected gene, which is expected due to the use of different screening criteria. A full list of genes used in NGSS1 and NGSS2, as well as their estimated linear coefficients in the models, are listed in Supplementary Tables E2 and E3.

Discussion

Transcriptomic analysis of host cells has proven informative in the study of several respiratory viral infections, including RSV.[15-18] Most reports have described gene

expression correlates of disease in peripheral blood mononuclear cells during infection since RSV pathogenesis is thought to be closely linked to the host's immune response.[26] In two publications from the same group, RSV infection was associated with overexpression of innate immunity genes (neutrophil and interferon genes) and suppression of adaptive T and B cell genes. [15, 17] The investigators used the results to develop a gene-expression based illness score (designated Molecular Distance to Health [MDTH]) that was significantly correlated with a clinical disease severity score, duration of hospitalization and need for supplemental oxygen. Similarly, we reported that gene expression patterns in purified CD4 T cells during infection were correlated with clinical disease severity.[16] In addition to circulating cells, gene expression results from nasal swabs during RSV infection has also been recently reported, with differentially expressed genes correlated with a clinical severity score.[18]

The use of nasal specimens in infants is attractive for a number of reasons. Nasal respiratory epithelial cells are the first cells infected and directly initiate early innate immune responses to RSV. The mucosa is also the site of migration of both innate and adaptive immune cells during and after infection. Importantly, we have shown that gene expression in nasal respiratory epithelial cells is highly concordant with published gene expression in lower respiratory tract epithelial cells, and thus should be a reasonable proxy for lung responses to RSV infection.[20] Of practical importance, collection of nasal epithelial cells is relatively non-invasive and simple to perform with minimal discomfort.

The ability to quantify RSV disease severity in young infants has been a frequent and important topic of discussion in the literature, and several approaches have been

proposed for evaluation and management of RSV-infected infants.[14] A variety of clinical parameters have been included in different severity scores, with incomplete agreement on the optimal factors to select. One reason is that many clinical signs of RSV infection in young infants, including cutaneous oxymetry measurements, can fluctuate frequently and rapidly during the course of illness, making assessment difficult. An objective biomarker reliably and continuously correlated with clinical severity could prove useful in determining if an infant could be safely discharged from the emergency department or a physician's office. It might also be valuable as an outcome measure in vaccine efficacy studies or for evaluation of prophylactic or therapeutic medicines.

In this report we describe the use of RNAseq-based analysis of gene expression data from nasal specimens, during and after RSV infection, to develop two nasal gene-expression severity scores (NGSS1 and NGSS2) that are highly correlated with a clinically derived global respiratory severity score (GRSS). We used marginal screening of all genes followed by PCA analysis and step-wise model selection to develop NGSS1, a multivariate linear classifier of severity. In cross-validation analysis, NGSS1 was strongly correlated with GRSS and was a fairly accurate classifier of binary disease severity. Furthermore, the score tracked well with the clinical improvement seen in subjects at day 28 of illness onset. Of particular note, we find that including uncorrelated supplementary genes enhances the prediction accuracy, and recommend this approach as a routine for future classification/prediction analyses based on high-throughput data with substantial correlation.

The 41 NGSS1 genes include cytokines (TNFSF10, IL6, and CXCL2), extracellular matrix proteins (VIM, MMP19, RPS15A, FKBP1A, and VCAN), inflammation regulators

(CXCL2, CD163), and components of various signaling processes (GNS, HAVCR2, PTPRC, CTSL, INHBA, IL6, MMP19, CXCL2, SLC39A8, CCDC80, VCAN, CD163). Many of these genes are only known to be involved in fundamental biological processes and are therefore novel in RSV research, including ST3GAL1 (a type II membrane protein) and ATP10B (ATPase Phospholipid Transporting 10B).

We noted that many of the differentially expressed genes associated with GRSS did not change between the acute and the convalescent time points. It is possible that these genes may simply be slow to return to baseline expression levels, in contrast to those genes selected for NGSS1. However, it occurred to us that “stable” genes could be predictive of severity regardless of when a nasal sample was obtained. We successfully identified a small set of stable genes whose expression patterns are capable of accurately classifying disease severity (NGSS2) beyond acute infection. The rationale for this approach was that these genes were less likely to represent illness-responsive genes and we speculated that they might prove valuable in developing a predictor that was also valid *prior to* illness onset. While NGSS2 is slightly less accurate than NGSS1 in predicting GRSS during acute illness, the association between NGSS2 and GRSS is still relatively strong. Although speculative, it is possible that determination of NGSS2 in a nasal sample prior to infection might predict disease severity if the infant was subsequently infected.

Interestingly, the 13 NGSS2 genes were broadly related to cytoplasmic activities (EXOSC10, PLK2, PPIC, CLDN10, MAP3K13, MT1G, PXN), ATP binding (SEPHS2) and phosphoprotein regulation (BCKDK, PLK2, MAP3K13); activities that may be less directly responsive to acute infection with RSV. These observations suggest that the best nasal transcriptome predictors of respiratory symptoms are not limited to those

genes that directly regulate the immune response to RSV infection. Many “down-stream” genes may be able to provide useful information that enhances the predictive power of well-known genes such as IL6 and CXCL2.

There are limitations to our study and conclusions. The number of subjects used is relatively moderate, and a larger confirmatory study will be necessary to validate these results. Secondly, although the NGSS1 declined for the severely ill infants when clinical symptoms had resolved, it would be useful to determine if NGSS1 tracked closely over the course of an illness. If so this would strengthen our conclusion that it may be a potential biomarker for clinical RSV disease severity. The results may not be valid for infants older than 10 months of age when infected with RSV, nor for infants with prematurity or other underlying medical conditions. Finally, speculation that NGSS2, which uses genes that do not vary between the acute illness visit and convalescence, might predict disease severity prior to infection demands careful prospective validation.

Tables and Figures

Table 1. Demographic data of subjects. *P*-values reported in the last column were either based on Fisher's exact test (if the variable is categorical) or Welch *t*-test (if the variable is continuous).

	mild (n=42)[a]		severe (n=64)[a]		p value
	n	Mean (SE) or %	n	Mean (SE) or %	
Global Severity Score	42	1.63(0.15)	64	6.13(0.22)	<0.001
Visit Age (months)	42	3.52(0.31)	64	3.24(0.3)	0.5122
Gestational Age (weeks)	42	39.05(0.19)	64	38.8(0.18)	0.3437
Birth Weight (kg)	42	3.32(0.11)	64	3.36(0.07)	0.7468
Family Size	42	4.43(0.44)	64	3.98(0.22)	0.3703
Days Since Disease Onset	42	4.31(0.27)	64	4.86(0.22)	0.1209
Breast Feeding Summary	42	1.56(0.19)	63	1.53(0.16)	0.8979
Sex					0.4275
Male	23	44.23	29	55.77	
Female	19	35.19	35	64.81	
Ethnicity					0.8018
Hispanic or Latino	8	42.11	11	57.89	
Non-Hispanic or Non-Latino	34	39.08	53	60.92	
Race					0.3115
Caucasian	23	37.1	39	62.9	
Other race	19	47.5	21	52.5	
Missing	-	0	4	100	
Delivery Type					0.3634
Vaginal	29	36.71	50	63.29	
C-section	13	48.15	14	51.85	
Smoking Exposure					1
True	14	38.89	22	61.11	
False	28	40	42	60	
RSV Strain					1
A	23	38.98	36	61.02	
B	18	39.13	28	60.87	
Missing	1	100	-	0	

[a] based on GRSS ≤3.5 (mild) or > 3.5 (severe)

Table 2. Performance of four models used in developing NGSS1 and NGSS2. Naïve and CV RSS are the mean residual sums of squares of the predictive model in the original and cross-validation analyses, respectively. Correlation are the Pearson correlation coefficient between the predicted severity scores and the clinically defined GRSS. Prediction accuracy is the percentage of correctly predicted mild (NGSS ≤3.5) or severe (NGSS >3.5) symptoms, compared with the same phenotype defined by the GRSS (mild: GRSS≤3.5; severe: GRSS>3.5).

	number of genes selected	Naïve RSS	Naïve Correlation	Naïve misclassified subjects (out of 106)	CV RSS	CV Correlation	CV prediction accuracy	CV misclassified subjects (out of 106)
Model 1	39 genes	1.234	0.909	15	2.743	0.797	77.4%	24
Model 2*	**41genes**	**0.884**	**0.935**	**9**	**2.681**	**0.813**	**89.6%**	**11**
Model 3	42 genes	0.920	0.933	13	2.119	0.844	78.3%	23
Model 4**	**13 genes**	**2.549**	**0.800**	**16**	**3.215**	**0.741**	**84.0%**	**17**

* designated NGSS1. ** designated NGSS2

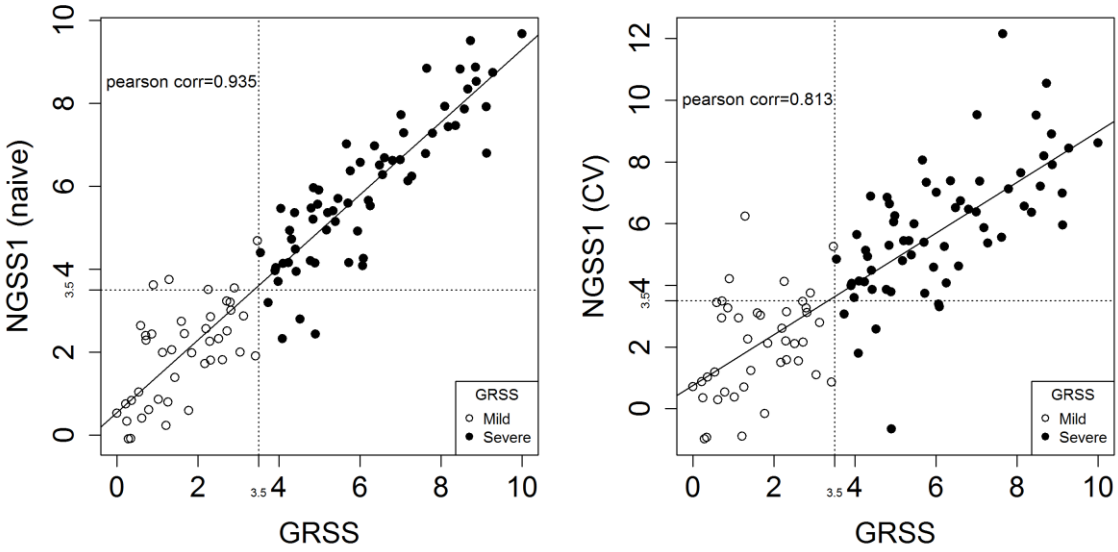


Figure 1. Correlating NGSS1 (severity score predicted by Model 2) with GRSS. Left: naïve Pearson correlation between GRSS and NGSS1 is $\boldsymbol{\rho = 0.935}$. Right: cross-validated Pearson correlation between GRSS and NGSS is $\boldsymbol{\rho = 0.813}$. Solid dots are subjects with severe symptoms (defined by GRSS>3.5) and empty dots are those with mild symptoms (GRSS≤3.5).

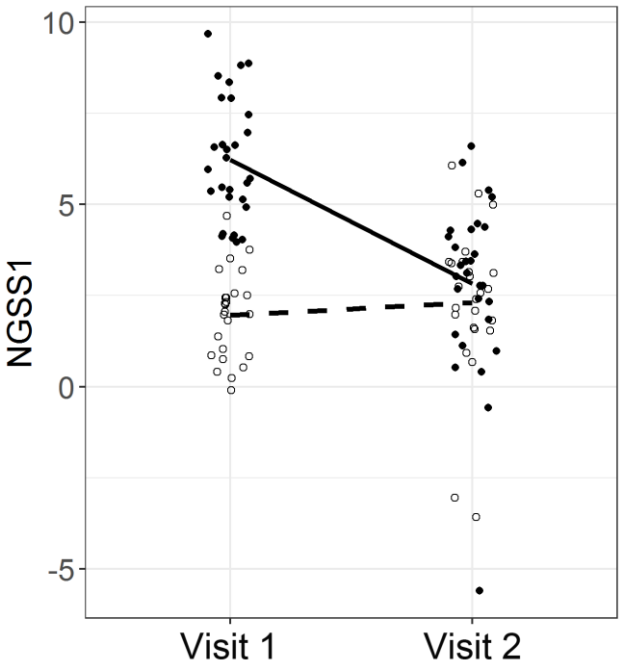
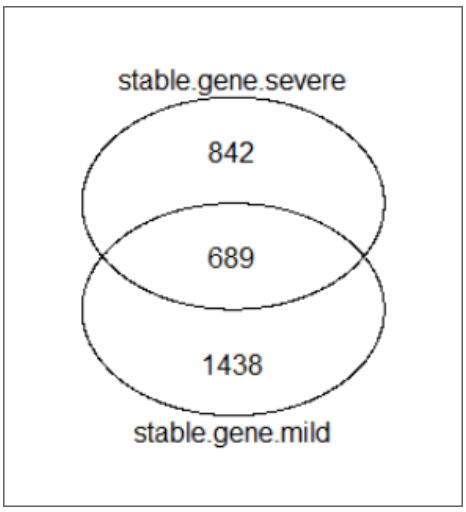


Figure 2. Paired comparisons between visit 1 and visit 2 using NGSS1. Solid dots represent severe subjects and empty dots represent mild subjects. The solid line represents the mean trend of severe subjects and the broken line represents the trend for mild subjects.

A.



B.

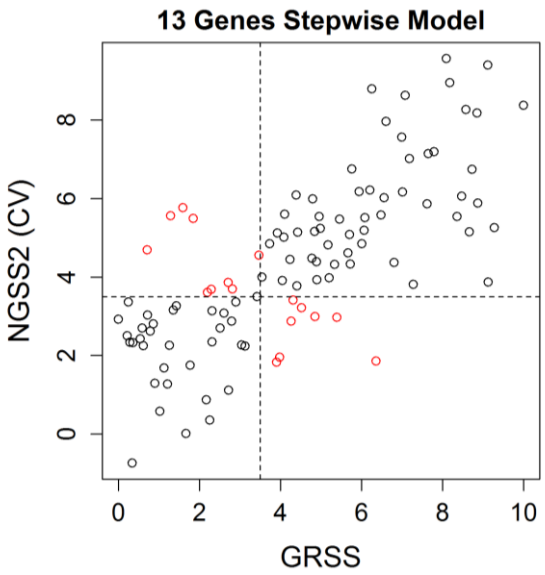


Figure 3. (A). Diagram indicating the stable genes for the mild (GRSS ≤3.5) and severe (GRSS >3.5) groups and the 689 intersecting stable genes common to both groups. (B). Correlating NGSS2 (severity score predicted by Model 4) with GRSS. Naïve Pearson correlation between GRSS and NGSS2 is $\boldsymbol{\rho = 0.800}$. Right: cross-validated Pearson correlation between GRSS and NGSS is $\boldsymbol{\rho = 0.741}$**.** Black circles are subjects with correct cross-validated classification based on NGSS2 and red circles are misclassified subjects.

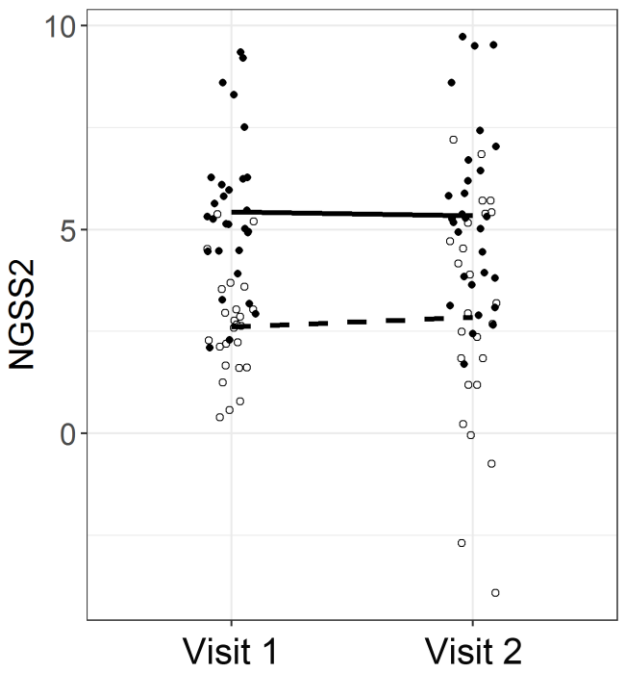


Figure 4. Paired comparisons between visit 1 and visit 2 for the 54 subjects with both samples using NGSS2. Solid dots represent severe subjects and empty dots represent mild subjects. The solid line represents the mean trend of severe subjects and the broken line represents the trend for mild subjects.

References

1. Hall CB, Weinberg GA, Iwane MK, Blumkin AK, Edwards KM, Staat MA, Auinger P, Griffin MR, Poehling KA, Erdman D *et al*: **The burden of respiratory syncytial virus infection in young children**. *N Engl J Med* 2009, **360**(6):588-598.
2. Shi T, McAllister DA, O'Brien KL, Simoes EAF, Madhi SA, Gessner BD, Polack FP, Balsells E, Acacio S, Aguayo C *et al*: **Global, regional, and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in young children in 2015: a systematic review and modelling study**. *Lancet* 2017, **390**(10098):946-958.
3. Hall CB, Weinberg GA, Blumkin AK, Edwards KM, Staat MA, Schultz AF, Poehling KA, Szilagyi PG, Griffin MR, Williams JV *et al*: **Respiratory syncytial virus-associated hospitalizations among children less than 24 months of age**. *Pediatrics* 2013, **132**(2):e341-348.
4. Bekhof J, Reimink R, Brand PL: **Systematic review: insufficient validation of clinical scores for the assessment of acute dyspnoea in wheezing children**. *Paediatr Respir Rev* 2014, **15**(1):98-112.
5. Corneli HM, Zorc JJ, Holubkov R, Bregstein JS, Brown KM, Mahajan P, Kuppermann N, Bronchiolitis Study Group for the Pediatric Emergency Care Applied Research N: **Bronchiolitis: clinical characteristics associated with hospitalization and length of stay**. *Pediatr Emerg Care* 2012, **28**(2):99-103.
6. Destino L, Weisgerber MC, Soung P, Bakalarski D, Yan K, Rehborg R, Wagner DR, Gorelick MH, Simpson P: **Validity of respiratory scores in bronchiolitis**. *Hosp Pediatr* 2012, **2**(4):202-209.
7. Duarte-Dorado DM, Madero-Orostegui DS, Rodriguez-Martinez CE, Nino G: **Validation of a scale to assess the severity of bronchiolitis in a population of hospitalized infants**. *J Asthma* 2013, **50**(10):1056-1061.
8. Feldman AS, Hartert TV, Gebretsadik T, Carroll KN, Minton PA, Woodward KB, Larkin EK, Miller EK, Valet RS: **Respiratory Severity Score Separates Upper Versus Lower Respiratory Tract Infections and Predicts Measures of Disease Severity**. *Pediatr Allergy Immunol Pulmonol* 2015, **28**(2):117-120.
9. Gajdos V, Beydon N, Bommenel L, Pellegrino B, de Pontual L, Bailleux S, Labrune P, Bouyer J: **Inter-observer agreement between physicians, nurses, and respiratory therapists for respiratory clinical evaluation in bronchiolitis**. *Pediatr Pulmonol* 2009, **44**(8):754-762.
10. McCallum GB, Morris PS, Wilson CC, Versteegh LA, Ward LM, Chatfield MD, Chang AB: **Severity scoring systems: are they internally valid, reliable and predictive of oxygen use in children with acute bronchiolitis?** *Pediatr Pulmonol* 2013, **48**(8):797-803.
11. Mosalli R, Abdul Moez AM, Janish M, Paes B: **Value of a risk scoring tool to predict respiratory syncytial virus disease severity and need for hospitalization in term infants**. *J Med Virol* 2015, **87**(8):1285-1291.
12. Parker MJ, Allen U, Stephens D, Lalani A, Schuh S: **Predictors of major intervention in infants with bronchiolitis**. *Pediatr Pulmonol* 2009, **44**(4):358-363.
13. Fernandes RM, Plint AC, Terwee CB, Sampaio C, Klassen TP, Offringa M, van der Lee JH: **Validity of bronchiolitis outcome measures**. *Pediatrics* 2015, **135**(6):e1399-1408.
14. Karron RA, Zar HJ: **Determining the outcomes of interventions to prevent respiratory syncytial virus disease in children: what to measure?** *The Lancet Respiratory medicine* 2018, **6**(1):65-74.
15. de Steenhuijsen Piters WA, Heinonen S, Hasrat R, Bunsow E, Smith B, Suarez-Arrabal MC, Chaussabel D, Cohen DM, Sanders EA, Ramilo O *et al*: **Nasopharyngeal Microbiota, Host Transcriptome, and Disease Severity in Children with Respiratory Syncytial Virus Infection**. *Am J Respir Crit Care Med* 2016, **194**(9):1104-1115.
16. Mariani TJ, Qiu X, Chu C, Wang L, Thakar J, Holden-Wiltse J, Corbett A, Topham DJ, Falsey AR, Caserta MT *et al*: **Association of Dynamic Changes in the CD4 T-Cell Transcriptome With Disease Severity During Primary Respiratory Syncytial Virus Infection in Young Infants**. *J Infect Dis* 2017, **216**(8):1027-1037.
17. Mejias A, Dimo B, Suarez NM, Garcia C, Suarez-Arrabal MC, Jartti T, Blankenship D, Jordan-Villegas A, Ardura MI, Xu Z *et al*: **Whole blood gene expression profiles to assess pathogenesis and disease severity in infants with respiratory syncytial virus infection**. *PLoS Med* 2013, **10**(11):e1001549.
18. Do LAH, Pellet J, van Doorn HR, Tran AT, Nguyen BH, Tran TTL, Tran QH, Vo QB, Tran Dac NA, Trinh HN *et al*: **Host Transcription Profile in Nasal Epithelium and Whole Blood of Hospitalized Children Under 2 Years of Age With Respiratory Syncytial Virus Infection**. *J Infect Dis* 2017, **217**(1):134-146.

19. Walsh EE, Mariani TJ, Chu C, Grier A, Gill SR, Qiu X, Wang L, Holden-Wiltse J, Corbett A, Thakar J *et al*: **Aims, study design and enrollment results from the Assessing Predictors of Infant Respiratory Syncytial Virus Effects and Severity (AsPIRES) study**. *Journal of Medical Internet Research* in press.
20. Chu CY, Qiu X, Wang L, Bhattacharya S, Lofthus G, Corbett A, Holden-Wiltse J, Grier A, Tesini B, Gill SR *et al*: **The Healthy Infant Nasal Transcriptome: A Benchmark Study**. *Sci Rep* 2016, **6**:33994.
21. Caserta MT, Qiu X, Tesini B, Wang L, Murphy A, Corbett A, Topham DJ, Falsey AR, Holden-Wiltse J, Walsh EE: **Development of a Global Respiratory Severity Score for Respiratory Syncytial Virus Infection in Infants**. *J Infect Dis* 2017, **215**(5):750-756.
22. Johnson WE, Li C, Rabinovic A: **Adjusting batch effects in microarray expression data using empirical Bayes methods**. *Biostatistics* 2007, **8**(1):118-127.
23. Zou H, Hastie T: **Regularization and variable selection via the elastic net (vol B 67, pg 301, 2005)**. *J R Stat Soc B* 2005, **67**:768-768.
24. Friedman J, Hastie T, Tibshirani R: **Regularization Paths for Generalized Linear Models via Coordinate Descent**. *J Stat Softw* 2010, **33**(1):1-22.
25. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: A practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society: Series B* 1995, **57**:289-300.
26. Collins PL, Fearns R, Graham BS: **Respiratory syncytial virus: virology, reverse genetics, and pathogenesis of disease**. *Curr Top Microbiol Immunol* 2013, **372**:3-38.